Use of simulation to investigate resourcing priorities and bed use in generic models of elective and emergency clinical pathways

Michael Allen, Steve Thornton and Matthew Cooke

Michael Allen, PhD*
Associate Professor of Clinical Systems Improvement
Warwick Medical School
Coventry
CV4 7AL

*Corresponding author and guarantor (e-mail: m.allen.1@warwick.ac.uk)

Professor Steve Thornton, DM FRCOG BM
Professor of Obstetrics and Gynaecology
Warwick Medical School
Coventry
CV4 7AL

Professor Matthew Cooke, PhD, MB, ChB, FRCS(Ed), FCEM, DipIMC
Director of Emergency Care and Systems Improvement Group
Warwick Medical School
Coventry
CV4 7AL

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd and its licensees, to permit this article (if accepted) to be published in BMJ editions and any other BMJPG products and to exploit all subsidiary rights, as set out in our licence (http://resources.bmj.com/bmj/authors/checklists-forms/licence-for-publication)"

"All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that (1) [initials of relevant authors] have support from [name of company] for the submitted work; (2) [initials of relevant authors] have [no or specified] relationships with [name of companies] that might have an interest in the submitted work in the previous 3 years; (3) their spouses, partners, or children have [specified] financial relationships that may be relevant to the submitted work; and (4) [initials of relevant authors] have no [or specified] non-financial interests that may be relevant to the submitted work."

Ethics approval was not required.

Use of simulation to investigate resourcing priorities and bed use in generic models of elective and emergency clinical pathways

Abstract

Objectives - to assess whether alternative methods of prioritising patients affects length of stay and bed use in simulation models of elective and emergency care pathways

Design - generic elective and emergency care pathways were modelled using process simulation software

Main outcome measures - length of stay, staff utilisation, bed occupancy

Results - Where *admission priority* (giving priority to bringing in new patients for start of treatment with priority reducing through to discharge) was used in a model of elective procedures length of stay continued to increase as bed numbers were increased despite the number of patients being treated and staff utilisation reaching a plateau at a lower bed number. Bed occupancy was consistently close to maximum even when an escalation or "unblocking" strategy was used to switch priority to the discharge step when there were few free beds available. Restricting bed numbers could avoid the increased length of stay. When *discharge priority* (always giving highest priority to discharge activities, with priority reducing back to admitting new patients) was used in the same elective surgery model length of stay was significantly reduced and length of stay and bed occupancy did not continue to rise as more beds were made available. When patient arrival was scheduled each day to match available clinical staff, application of *discharge priority* reduced length of stay and bed occupancy by about a third compared to *admission priority*. In an emergency care setting (where there is no control over patient arrival) length of stay within the emergency department increased as patient arrival rate increased with large increases in waiting time observed above 80% capacity utilisation. Application of *discharge priority* (for non-urgent cases) reduced average length of stay by a third or more compared to *admission priority* at high capacity utilisation.

Conclusions – the modelling suggests that the length of stay in elective wards or emergency departments and bed occupancy in elective wards may be significantly reduced by, in the absence of other urgent medical need, constantly giving highest priority to discharge activities, with reducing priority back through the care pathway. An escalation strategy of bed unblocking (prioritising discharge activities only when bed occupancy is close to maximum) may have little impact on overall length of stay.

1. Introduction

Healthcare systems are frequently under pressure to maximise the number of patients treated (and maximise use of staff and other costly resources) while minimising patient waiting times and length of stay in hospital in order to increase patient satisfaction[1], reduce risks of hospital-acquired infections[2] and reduce hospital costs[3]. There have been many initiatives around reducing lengths of stay and improving flow through the hospital. These have often focussed on timely discharge[4] and bed management[5] and the need to reduce bed occupancy, which has been shown to be associated with increased availability of beds for emergencies[6] and decreased waits in the emergency department[7]. Many methodologies have been used for improving the performance of the system, including the use of Lean techniques[8], Six Sigma[9], manufacturing- like Material Requirements Planning (MRP) systems [10] or other informatics systems[11]. Discrete event stimulation has also been widely used as a tactical tool to aid the optimisation of specific departments such as emergency departments [12,13], operating rooms[14], orthopaedics[15] and radiography[16]. However generalised models have not been used to aid the understanding of hospital system behaviours at a higher hospital-wide level.

In this paper we look at two model systems which represent the two basic models of hospital treatment. The first model is an elective treatment model where it is assumed that there is a pool of patients in the community on a waiting list for treatment. The waiting list allows the hospital to govern its admissions and only "pull" patients in at a rate that matches its ability to process the patients through the hospital. The second model represents an on-demand system where patients turn up at the hospital at will; this model mimics the behaviour of emergency departments or "walk in" clinics. The hospital has little or no control over patient arrival except perhaps in extreme cases such as closing the emergency department to minor injuries if overwhelmed by a major disaster.

We use the generic elective treatment model to examine the relationship between bed availability, internal hospital staff availability, length of stay and the number of patients treated. We examine two alternative patient prioritisation strategies; the first prioritises bringing patients into available beds and starting treatment ("*admission priority*") and the second prioritises later stage process and discharge activities ("*discharge priority*"). The prioritisation rules do not apply just to single steps, but flow through the whole system, so that in *discharge priority* staff move to the latest stage of work/treatment that is waiting, whereas in *admission priority* staff move to the earliest stage work/treatment that is waiting. We also examine the impact of a bed "unblocking" strategy where priority is switched to discharge activities only when bed occupancy is near full, which is common practice in the NHS in response to limited free bed availability (through escalation plans).

In the on-demand model we look at prioritisation so that after an initial triage, which highlights urgent cases which always have highest priority, staff may prioritise the earliest waiting non-urgent patient ("*admission priority*") or may prioritise the latest stage of treatment ("*discharge priority*").

2. Method

2.1. Elective treatment model

The elective treatment model is shown in figure 1.

In the elective model the doctor first permits the acceptance of a new patient into the system. The patient is then admitted by a nurse. The doctor sees the patient before the surgery or (or other procedure) and is then responsible for the procedure. The patient recovers for 24 hours before being assessed and discharged by the doctor.

All timings except recovery have 100%CV (where standard deviation = mean) based on a log-normal distribution (the log normal distribution is typical of manual operations and is used to describe processes where a few jobs take significantly longer than the average time, creating a right-skewed distribution). Recovery is a fixed period of 1 day. The model is constructed so that the doctor is the constraining human resource (the doctor is the busiest person and limits the number of patients that may be treated when free beds are available). The number of beds is varied as one of the parameters in the model. In this elective model it is assumed that there is an inexhaustible pool of patients to treat and so the only time measured in the model is the length of stay within the hospital. The model has two doctors and two nurses.

A modification to the model is also described where, rather than pulling from an unlimited pool of patients, 12 new patients are scheduled to arrive each morning; this number of patients is set so that on average 95% of the doctors' time is expected to be used each day.

A second modification to the model is where the nurse (who is not a constraining resource in this system) is allowed to permit the admission of new patients into the system whenever there is a bed available.

2.2.On demand (emergency) clinic

The on-demand clinic model is shown in figure 2.

In the on-demand clinic model patients arrive at random (with an exponential distribution of inter-arrival times, which is the distribution expected for random arrivals) during the day. The arrival pattern is observed as being random; it is quite possible for several patients to arrive close together followed by a long period with no arrivals. The average inter-arrival time is varied in the model to examine the impact of staff utilisation (as the average inter-arrival time is reduced there is an increased number of arrivals leading to higher staff utilisation).

Patients are admitted by reception and then see a nurse. They are seen by a doctor before having an X-ray or other imaging. They are then seen by the doctor again before being treated by a nurse and discharged.

The model has two doctors, three nurses and two imagers always on duty. All timings have 100%CV (standard deviation = mean) based on a log-normal distribution. The model is constructed so that the doctor is the constraining resource.

10% of patients are randomly assigned as being urgent in which case they always have highest call on resources, overriding other prioritisation rules in place.

2.3.Prioritisation strategies

Two prioritisation strategies are assessed in the model:

Admission priority: unless overridden by medical priority (urgent cases) staff move to examine or treat waiting patients at the earliest step of the process (admission procedures take priority over discharge procedures; starting the treatment process for new patients takes priority over completing the treatment process for patients already in progress). Priority decreases as the patient moves through the system.

When using *admission priority* an option was available for "unblocking" beds. When this option is selected priority is shifted to discharge when there is one bed or no bed available for new patients. When there are two or more free beds available priority returns to the normal rules.

Discharge priority: unless overridden by medical priority (urgent cases) staff move to examine or treat patients waiting at the latest step of the process (discharge procedures take priority over admission procedures; completing the treatment process for patients already in progress takes priority over starting the treatment process for new patients). Priority increases as the patient moves through the system.

2.4. Technical information

The model was built in Simul8 2009 (SIMUL8 Corporation). All results are the average result of multiple runs of the model using a different random number seed for each run. The number of runs (120 in both models) per scenario was set to give 95% confidence limits of no greater than 5% of the mean results. All models had an equilibration period of 50 days (which primes the system, starting from an initially empty system) followed by data collection for 100 days.

3. Results

3.1. Elective treatment model

The number of patients treated initially increased with bed availability, but as staff become more highly utilised increasing the number of beds had reduced impact and then had no further impact on the number of patients treated (Fig. 3A) due to maximum utilisation of the busiest staff (the doctors). When measuring the number of patients treated, *admission priority* (with or without unblocking) and *discharge priority* behaved very similarly, with a similar maximum number of patients treated and a similar relationship between the total number of beds available and patients treated.

The minimum length of stay (1.2 days) was achieved when the number of available beds was most limited (corresponding to lowest doctor utilisation). This minimum length of stay reflects the sum of the process step times, with no unnecessary waiting between the steps. Average length of stay increased as the number of available beds increased (Fig. 3B), but the impact was greater when *admission priority* was applied (length of stay increased in proportion to the number of available beds) than when *discharge priority* was used (average length of stay plateaued at 1.7 days). At 40 beds, for example, average length of stay per patient was twice as long when *admission priority* was applied compared to *discharge priority* (3.2 days *c.f.* 1.6 days) despite the number of patients treated remaining the same. Applying an escalation unblocking strategy to *admission priority* (giving highest priority to discharge when less than one free bed is available) did not improve length of stay results for *admission priority*, though the model revealed that the queue moved from being predominantly before discharge to predominantly before the procedure step. For example in a 40

bed system (Fig. 4) applying *admission priority* patients waited an average of 0.1 days before the procedure, but then waited 2.0 days for discharge. When unblocking was used the wait for discharge reduced to an average 0.1 days, but the queue for the procedure rose to 1.9 days. Discharge priority reduced queuing at all steps apart from the first step (first diagnostic step), with total unnecessary waiting (queuing) time in the system reduced from 2.1 days to 0.6 days.

Bed occupancy rose with length of stay (Fig. 3C); when *admission priority* was applied bed occupancy rose and stayed very close to the total number of beds available (Fig. 3D), but with *discharge priority* bed occupancy rose (to about 20 beds) and then reached a plateau despite free beds being readily available. Applying unblocking to *admission priority* had only minimal effect on bed occupancy, for example in a 40 bed system, bed occupancy averaged 99% with or without unblocking (release of beds was very quickly followed by re-occupation of those newly available beds). Unblocking may appear to be having an effect on the system, as the queue for discharge reduced dramatically. However the queue simply moved from before discharge to before the procedure (Fig. 4).

When patient arrival was fixed at 12 patients per day (arriving at the start of the day) in a 40 bed system, discharge priority reduced average length of stay from 2.2 to 1.6 days and reduced average bed occupancy from 26 to 17 patients (Fig 5).

In the results above patient entry into the system was controlled by the doctor (the limiting resource). Figure 6 shows results obtained if a non-limiting resource (nurse) permits entry of patients into the system. When there was no limit to the potential number of patients arriving in the model it was important that the limiting resource (the bottleneck resource, or the busiest type of staff, that limits the throughput of the system) controlled the admission of new patients in the *discharge priority* method and that admittance of new patients was the lowest priority task they performed. If the limiting resource did not control patient admission then bed occupancy again naturally rose to near full bed occupancy with length of stay rising in proportion to bed occupancy. If patients were scheduled to arrive at a rate within the capacity of the limiting resource then there was no need for the limiting resource to be in control of patient admittance.

3.2.On-demand model

In the on-demand model length of stay increased as the arrival rate of patients increased (Fig . 7A). Minimum length of stay (0.6 hours) was achieved when patient arrival rate was at its lowest. Average length of stay doubled when patients were arriving at ~75% system capacity. As utilisation of system capacity increased above 75% the deterioration in length of stay was more marked when *admission priority* was used compared to *discharge priority* so that at 90% system capacity average length of stay using *admission priority* was 2.8 hours compared to 1.8 hours using *discharge priority*. Correspondingly there is a difference in the number of patients completed within 4 hours (Fig. 7B); at 90% system capacity 78% of patients had been discharged within 4 hours under *admission priority* rules, compared to 96% patients when using *discharge priority* rules. The number of patients within the unit increased in proportion to the length of stay and increased non-linearly with system utilisation (Fig. 8). The number of patients in the unit was greater when using *admission priority* rules compared with *discharge priority* rules, so that at 90% system capacity there was an average of 25 patients in the unit using *admission priority* rules compared to 16 patients using *discharge priority* rules.

The difference in length of stay (and number of patients in the unit) between *admission priority* and *discharge priority* rules increased as the unit became busier (Fig 9). At 80% capacity utilisation average length of stay was 28% longer using *admission priority* rules, but at 90% capacity utilisation was 56% longer.

The variability in length of stay was approximately proportional to the average length of stay, with the coefficient of variation (standard deviation/mean) ranging from 45% at the staff recourse utilisation to 60% at the highest staff utilisation (data not shown). Thus as length of stay increases the unpredictability in the system also increases.

4. Discussion

4.1. Elective treatment model

The first model of an elective treatment process demonstrated how applying staff with highest priority to discharge activities and then reducing priority back to admission activities leads to a reduced length of stay that is associated with reduced bed occupancy. In the opposite method of prioritising staff (giving highest priority to admitting new patients when a bed becomes free) bed occupancy always rises to close to maximal even when the true constraint is elsewhere (staff become close to maximally utilised when 20 beds are present in the model and yet bed occupancy is maintained at close to 100% no matter how many beds are available).

If admissions are governed by available beds (bringing in patients when there is a bed available) then an observer may view the high bed occupancy (being close to 100%) as a key factor in constraining the number of patients who may be treated, but the modelling shows that a temptation to increase the number of available beds could lead to increased length of stay (and therefore increased costs) without increasing the number of patients treated and without reducing percentage bed occupancy significantly below 100%. This supports previous observations which showed no clear correlation between number of beds and the number of patients treated[17].

Unblocking activities (giving priority to discharge only when bed occupancy is full or close to full) had little impact of the performance when *admission priority* rules were followed. Though the number of patients waiting for discharge was dramatically reduced, the queue simply moved to within the system to an earlier step (before the operation) with no overall change in length of stay and no overall reduction in patients waiting for an activity. These results highlight a danger of focussing just on “blocked bed days” (days where a patient remains though they are ready for discharge), as purely focussing on eliminating these blocked bed days may simply move the queue to elsewhere within the system where it becomes hidden, with no overall reduction in length of stay. Although a push to discharge patients when capacity is full is common practice in the NHS it is unlikely to solve any problems in the medium to long term and has minimal impact on average length of stay in the model.

In a *discharge priority* system staff are constantly prioritised to discharge even when empty beds are available. Prioritisation then flows back down the system with new elective admission the lowest priority activity. Prioritising discharge activities even when there are many empty beds available and patients are waiting to enter the system may seem counter-intuitive, as might preventing or delaying the admission of patients when bed occupancy is very low, but the model predicts that prioritisation around the whole flow may substantially reduce length of stay and reduce bed occupancy.

These results are consistent with Little's law [18] where the average length of stay will equal the average number of patients within the system divided by the throughput of patients (patients treated per unit time). When throughput is limited by hospital staff and not the number of beds, applying unblocking activities will not reduce length of stay unless a significant reduction in bed occupancy is also seen; giving priority to unblocking beds only to immediately admit a new patient will not reduce the number of patients in the system and so will not be expected to result in reduced length of stay, according to Little's law. When patient throughput is limited by staff, and not by bed number, reducing number of available beds may actually enhance the performance of the system, reducing length of stay (and therefore costs) without significantly effecting the number of patients treated (or care of the patient). Though this may seem counter-intuitive, this is analogous to the production control method known as Constrained Work In Progress or ConWIP where manufacturing cycletime (the time from start of manufacture through to completion) is controlled by restricting the quantity of inventory allowed in a production line [19]. We may see the number of beds as controlling the work in progress (patients) within the system. Initially as work in progress (patients) in the system increases the number of patients treated also increases as staff become more heavily utilised. However increasing work in progress beyond a critical level ("critical WIP") leads to diminishing returns in throughput gains but ever increasing length of stay within the system. The *admission priority* system described here behaves like a ConWIP system – pulling in more work at the front end until a predetermined limit (the ConWIP limit, or number of beds in our system is reached). In ConWIP systems setting the desired level of WIP is key. If the number of beds is too low then staff are under-utilised and patient throughput is limited. However if number of beds is too high then unnecessary queues build up in the system and length of stay becomes elongated without any increase in the number of patients treated. With high ConWIP levels (high number of beds), even with an unblocking strategy the system remains full as work is always pulled in at the front end when the capacity of the system (beds) allows. A low number of beds leads to a situation where staff are waiting for patients to treat, whereas a high number of beds leads to a situation where patients are waiting for staff.

With *discharge priority* the work in progress (patients within the system) finds its own minimal level while maximising the utilisation of the staff. If *discharge priority* (with priority flowing down through the system and the limiting resource controlling entry into the system) is employed and bed occupancy is still 100% then the constraint is very likely to be the bed number rather than staff and increasing the bed number should increase patient throughput without leading to a significant increase in length of stay.

The increased bed occupancy expected with *admission priority* or with escalation bed unblocking strategies may also have a significant impact on emergency department length of stay; hospital bed occupancy of greater than 90% has been associated with increased length of stay in the emergency department due to delays in accepting patients from ED into the general hospital system[20]. A constant prioritisation of discharge in the main hospital should therefore benefit the ED length of stay.

In scenarios where beds were the true constraining factor (when no staff were working at 100% utilisation) neither the choice of prioritisation rules nor unblocking had significant impact. For example in a 10 bed system where utilisation elsewhere was 85% or less the average length of stay ranged from 1.19-1.22 days between the three prioritisation methods. When beds are the true

constraining resource prioritising other resources (staff) has minimal impact as they are not constraining the system performance. In hospitals where beds rather than internal resources are limiting, staff have the time to complete all required tasks on time. The model predicts that it is when staff (rather than bed availability) are limiting the system that queuing within the system develops and prioritisation of work for those staff makes a significant difference to length of stay.

The published evidence on the impact of changing bed availability on length of stay (and associated costs) is mixed. In an international survey of acute hospital performance de Looper and Bhatia [21] concluded that there is an apparent correlation between availability of beds and length of stay, the higher the bed density per thousand population, the longer the hospital stay. Though we cannot say that our model explains such a correlation (as the reasons are likely to be multifactoral) the observed correlation is consistent with the behaviour of a system that is constrained by hospital staff rather than by available beds and where admission is governed by the availability of beds. The reduction of length of stay with reducing available beds has been noted by others [22] but was not observed in another survey of NHS resources[23], and reducing bed availability has also been associated with an increase in length of stay[24]. It has been suggested that limiting available beds may simply lead to increased pressure to improve services (or to discharge patients before they would otherwise be discharged) , and thereby cause an indirect reduction in length of stay [23]. Though that is possible, or even likely, this model shows that even without process improvement (and without reducing time spent on active care) limiting bed availability may be expected to reduce length of stay simply by controlling work in progress. It should be emphasised that in this model changing bed availability was not accompanied by changes to staff levels or by discharging patients before they are clinically ready. Data available from the literature that examines the impact of bed closures is generally confounded by simultaneous changes in the availability of other hospital resources or staff and may also reflect a required change in clinical practice.

When using *discharge priority* to control patient intake it is important to use the limiting resource to be the control step for admitting new patients – hence in the model the doctor (who is the limiting resource) permits the intake of new patients and they only do that when no other work is available. This is similar to the Theory of Constraints production control method known as drum-buffer-rope where new work is allowed into a bottlenecked production facility only as the bottleneck releases work [25].

Discharge priority can be expected to significantly reduce length of stay when the constraint in a system are staff that work at more than one point in the system. *Discharge priority* would not be expected to significantly impact length of stay where the constraint is a resource that is only ever used once during the patient stay. Where the constraint is a single step in the process any queuing builds up before that point and then work (or a patient) flows freely through the remainder of the process as all other resources have spare capacity. This may occur if throughput is limited by a specialised single piece of equipment (such as a scanner or access to an operating theatre) rather than being limited by staff or beds across the care pathway.

It should be noted that reducing length of stay does not necessarily lead to increased throughput in the system. When staff are maximally utilised throughput is not increased by changing choice of priorities, despite this prioritisation having a significant effect on length of stay.

In the model we also ran a scenario where patients were scheduled to arrive each morning to match the known capacity of the clinical staff (avoiding the temptation of allowing more patients into the system even when beds are available). This model more accurately describes an isolated system where the capacity constraint (doctors in this case) is known and patients may be scheduled to arrive to match the staff availability. In this scenario applying discharge priority reduced average length of and bed occupancy by about a third.

The differences between *admission priority* and *discharge priority* may actually be greater in real life as an extension of lengths of stay may increase staff utilisation as they continue to monitor and care for the increased number of patients; this increase in staff utilisation will very likely exacerbate queuing and congestion within the system.

4.2.On-demand model

The on-demand model is simpler than the elective model as there is no control over patient arrival or bed number.

In the model of the "on demand" clinic (which may mimic situations such as emergency departments and "walk in" clinics) the length of stay in the system increased as the patient load increased. This relationship is well established in queuing theory with a non-linear relationship between staff utilisation and waiting times[19] and large increases in length of stay above 80% staff utilisation. At high staff utilisation the system also becomes very sensitive to small changes in demand – small increases in demand can lead to very significant changes in waiting times and in the percentage of people treated within a defined period. At high staff utilisation the variability in length of stay is proportional to the average length of stay, and so the variability in the system increases as the system gets busier. This is a challenge when trying to combine high staff utilisation with average length of stay targets as the system may easily switch between flowing and congested with little apparent cause.

Again, even though we do not control patient arrival in the on-demand model, length of stay was reduced by applying *discharge priority* rules. The effect of using *discharge priority* over *admission priority*, after an initial triage, increased with staff utilisation. As with the elective admission model these savings in length of stay in the model are achieved by reducing queuing within the system rather than by changing any of the process activities within the system. Length of stay for high priority (urgent) patients was unaffected by the choice between *discharge priority* and *admission priority* as these patients always were the first to call on staff regardless of their position in the system.

The increased congestion associated with an *admission priority* system may have an effect on patient care and safety. Sprivulis *et al*. [26] have demonstrated that hospital and ED overcrowding is associated with higher mortality rates.

4.3.Application in clinical work

In the elective care model the best flow was achieved by always giving priority to the discharge process. Unblocking strategies to fire fight the non-availability of beds does not improve throughput or average length of stay but merely changed the location of the queue. This style of work can be directly applied to elective units where the clinical plan of each pathway will be decided before arrival of the patient and so does not need an early doctor intervention. This model suggests that

doctors (and supporting staff and utilities) in elective units should start each day by discharging patients and continue to give this work priority throughout each day.

The model of emergency care can be used to apply to either flow through an emergency department or the emergency flow through a whole hospital. Once again it shows that the best flow is achieved in the model by giving discharge a high priority and that fire fighting discharges is not a productive activity. The model does however create some clinical conflict. A recent NCEPOD report has shown that early review by a doctor is associated with lower rate of early deaths in hospital[27]. There is a complex interaction of factors that may improve safety that includes both early review (*admission priority*) but also involves increased availability of doctors to spend more time on complex cases. One model may be to have emergency or acute medicine doctors seeing all patients soon after arrival (in order to triage patients and identify those requiring urgent treatment) and other doctors in the hospital adopting a discharge priority approach to their usual activity unless urgent treatment is prescribed for a patient.

In this paper we have investigated the general principles involved in bed management and resource prioritisation and not dealt with the practical implementation of how a *discharge priority* system (that goes beyond simply prioritising the discharge step to unblock beds) might be deployed. Though these are simplified models they demonstrate underlying system behaviour as the number of available beds change and the prioritisation of activities are changed. Future research will focus on possible practical methods of implementation of a discharge priority system in a live hospital setting whilst maintaining early doctor review of emergency cases to maintain clinical safety.

Below are two cases demonstrating how this modelling may translate into clinical practice.

Case Study one:

An acute trust has seen that over the last two years there has been a gradual increase in the length of stay of patients. Investigation revealed that there had been no change in the case mix being seen. Escalation plans had been developed including undertaking extra discharge ward rounds once a certain level of occupancy or decreased flow was recognised but the situation only seemed to improve temporarily. Staff felt they were working at maximum capacity. There were calls for more beds to increase capacity and improve flow.

What this study suggests:

1. Flow may improve by changing from a conditional discharge focus to a discharge focussed approach and this may decrease length of stay, causing a decrease in workload and some decrease in staff utilisation

2. It may be more appropriate to focus on decreasing staff utilisation than on increasing bed numbers (this could be more staff or could be decreasing workload of high utilisation staff by reapportioning work or redesigning systems e.g. decreasing duplication).

Case Study two:

Flow through the emergency department appears slow and often the department is overcrowded with many more patients than there are cubicles. This seems to occur even on days when there are

beds available in the hospital and hence there is no exit block. There was nearly always a queue of patients to be seen so staff felt they were being 100% utilised. One major resuscitation case would cause delays for other patients for several hours. Internal measures suggest that there is often a long delay to start assessment of patients, it was calculated that most patients left the department within 3 hours of being first assessed. Therefore an internal target of seeing patients within one hour of arrival has been set and the focus is on making sure the queue in the waiting room is reduced.

What this study suggests:

1. Focusing on initial assessment may mean there are more patients in the department at any time and that length of stay increase, throughput and length of stay may be further worsened by the increased workload of monitoring patients and the effect of having no cubicles available.

2. Focussing on discharging patients will mean fewer patients in the department and better throughput hence reducing the 3 hours from assessment time to discharge (the model still gives urgent time-critical cases priority to be seen and the discharge focus is for the non-time critical cases).

3. High utilisation rates mean that there was little resilience in the system for the random arrival of a serious cases requiring significant manpower and decreasing utilisation rates would both improve flow as well as create a system more resilient to fluctuations in workload.

4.4.Summary

In summary, the modelling suggests that the length of stay in surgery (or medical) wards or emergency departments and bed occupancy in surgery wards may be significantly reduced by, in the absence of other urgent medical need, constantly giving highest priority to discharge activities, with reducing priority back through the care pathway. Bed unblocking strategies (prioritising discharge activities only when bed occupancy is close to maximum) may have little impact on overall length of stay.

References

1. Wigder HN, Johnson C, Shah M, Fantus R, Brasic J, Tanouye K, et al. Length of stay predicts patient and family satisfaction with trauma center services. *American Journal of Emergency Medicine* 2003;21:606-07.
2. Tess BH, Glenister HM, Rodrigues LC, Wagner MB. Incidence of hospital-acquired infection and length of hospital stay. *European Journal of Clinical Microbiology & Infectious Diseases* 1993;12:81-86.
3. Kaboli PJ, Barnett MJ, Rosenthal GE. Associations With Reduced Length of Stay and Costs on an Academic Hospitalist Service. *Americal Journal of Managed Care* 2004;10:561-68.
4. DH. Achieving timely 'simple' discharge from hospital - a toolkit for the multi-disciplinary team: Department of Health: London., 2004.
5. DH. Bed management demand and discharge predictors: Department of Health, London, 2004.
6. Bagust A, Place M, Posnett JW. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *British Medical Journal* 1999;319:155-58.
7. Cooke MW, Wilson S, Halsall J, Roalfe A. Total time in English accident and emergency departments is related to bed occupancy. *Emergency Medicine* 2004;21:575–76.
8. Brandao de Souza L. Trends and approaches in lean healthcare. *Leadership in Health Services* 2009;22(2):121-39.
9. Feng Q, Manuel CM. Under the knife: a national survey of six sigma programs in US healthcare organizations. *International Journal of Health Care Quality Assurance* 2008;21:535 - 47
10. Roth AV, Van Diredonick R. Hospital resource planning: concepts, feasability and framework. *Production and operations management* 1994;4(1):2-29.
11. Proudlove NC, Boaden R. Using operational information and information systems to improve in-patient flow in hospitals. *Journal of Health Organisation and Management* 2005;19:466-77
12. van Oostrum JM, Van Houdenhoven M, Vrielink MM, Klein J, Hans EW, Klimek M, et al. A simulation model for determining the optimal size of emergency teams on call in the operating room at night. *Anesth Analg* 2008;107:1655-6.
13. Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, et al. Forecasting emergency department crowding: a discrete event simulation. *Ann Emerg Med* 2008;52:116-25.
14. Marjamaa RA, Torkki PM, Hirvensalo EJ, Kirvela OA. What is the best workflow for an operating room? A simulation study of five scenarios. *Health Care Manag Sci* 2009;12:142-6.
15. Cipriano LE, Chesworth BM, Anderson CK, Zaric GS. An evaluation of strategies to reduce waiting times for total joint replacement in Ontario. *Med Care* 2008;46:1177-83.
16. Coelli FC, Ferreira RB, Almeida RM, Pereira WC. Computer simulation and discrete-event models in the analysis of a mammography clinic patient flow. *Comput Methods Programs Biomed* 2007;87:201-7.
17. Fowkes FGR, Page SM, Phillips-Miles D. Surgical manpower, beds and output in the NHS: 1967-1977. *British Journal of Surgery* 2005;70:114-16.
18. Little JDC. A Proof of the Queueing Formula L = λ W. *Operations Research* 1961;9:383-87.
19. Hopp W, Spearman M. *Factory Physics*. 2nd ed: McGraw-Hill/Irwin, 2000.
20. Forster AJ, Stiell I, Wells G, Lee AJ, van Walraven C. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad Emerg Med* 2003;10(2):127-33.
21. de Looper M, Bhatia K. *Australian Institute of Health and Welfare 1998. International health - how Australia compares.*: Australian Institute of Health and Welfare, Canberra (AIHW cat. no. PHE 8), 1998.
22. Marshall RD, R.I. S. A more efficient use of hospital beds? *British Medical Journal* 1974;3:27-30.
23. Martin S, Smith P. Explaining variations in inpatient length of stay in the National Health Service. *Journal of Health Economics 15* 1996;15:279-304.
24. Shepard DS. Estimating the effect of hospital closure on area wide inpatient hospital costs: a preliminary model and application. *Health Serv Res* 1983;18:513–49.

25. Goldratt EM, Cox J. *The Goal: A Process of Ongoing Improvement* North River Press, 1992.
26. Sprivulis PC, Da Silva J-A, Jacobs IG, Frazer ARL, Jelinek GA. The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments.[Erratum appears in Med J Aust. 2006 Jun 19;184(12):616]. *Med J Aust* 2006;184(5):208-12.
27. Cooper H, Findlay G, Goodwin APL, Gough MJ, Lucas SB, Mason DG, et al. *Caring to the End? A review of the care of patients who died in hospital within four days of admission*: National Confidential Enquiry into Patient Outcome and Death, London, 2009.

Figures

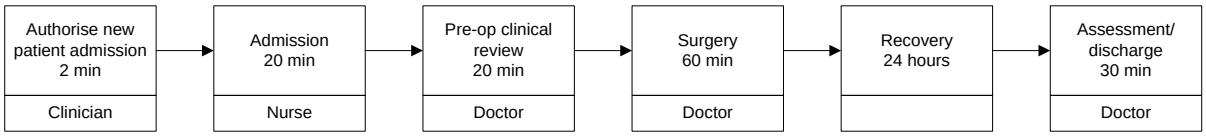

Fig.1. Elective treatment process model.

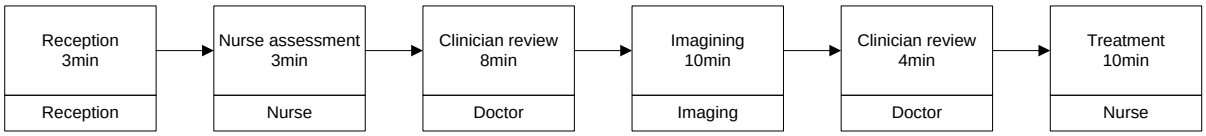

Fig.2. On-demand clinic model.

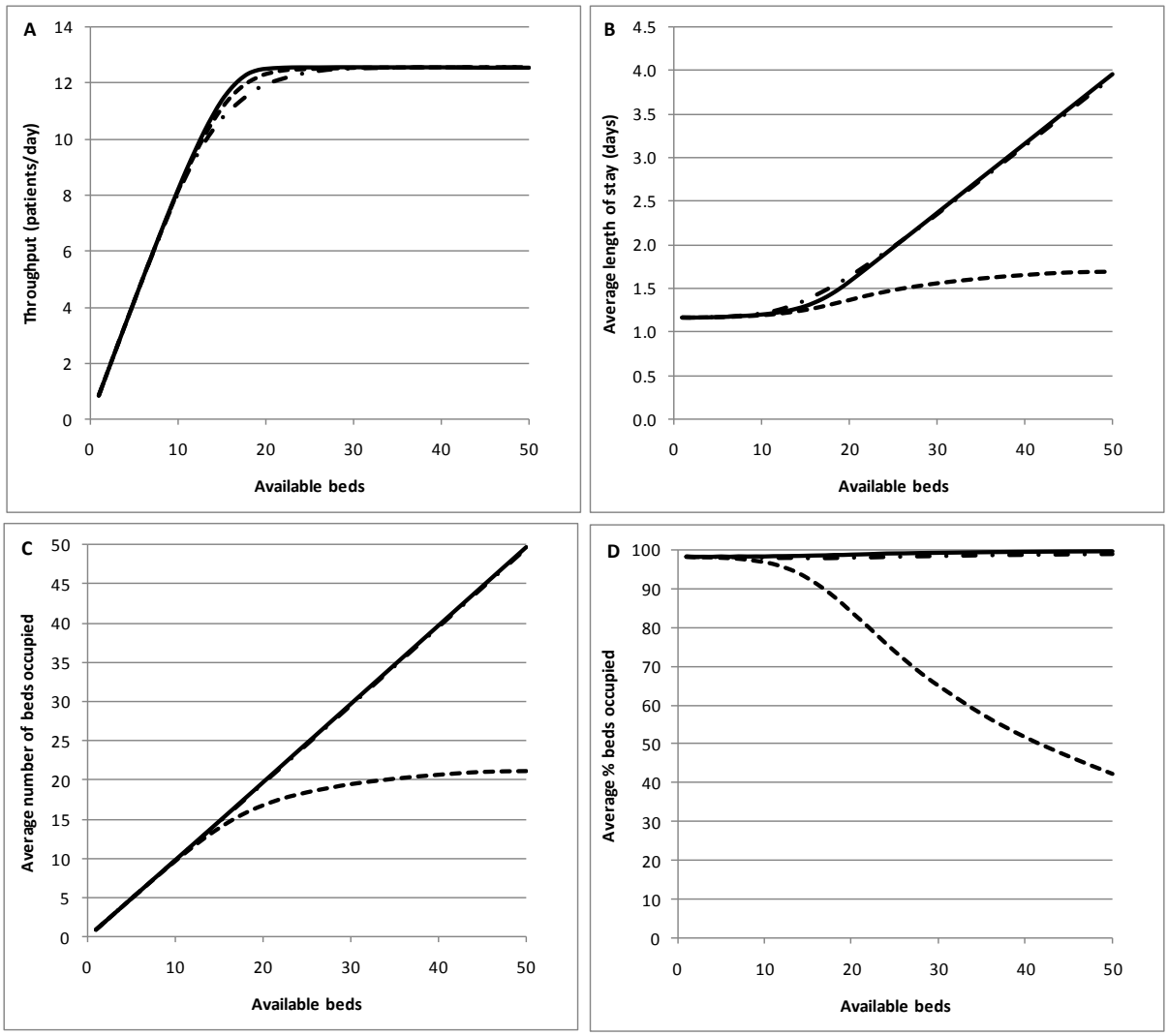

Fig. 3. The impact of changing the number of available beds on (A) patient throughput, (B) average length of stay, (C) average number of occupied beds, and (D) percentage bed occupancy. The solid line shows results when dispatch priority is used, the dotted line when *admission priority* is used and the dash-dot lined when *admission priority* is combined with unblocking.

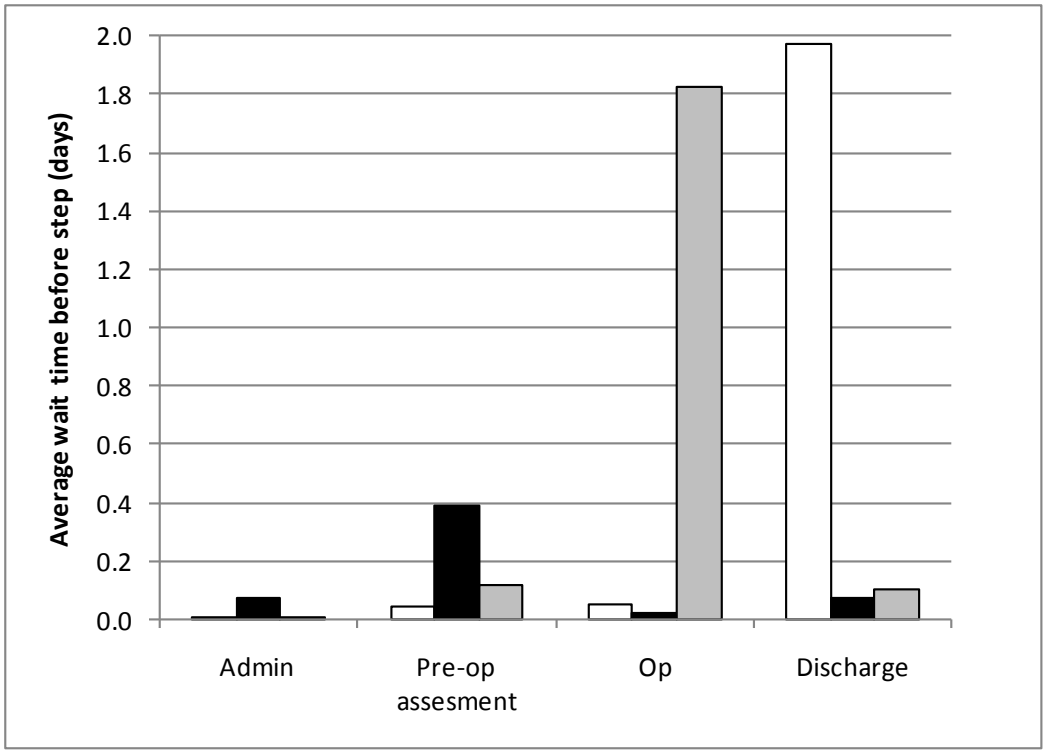

Fig. 4. The intra-hospital waiting times with different prioritisation rules (in a 40 bed model). The waiting times shown are average waiting times from when a patient is ready to move to the next step. Open bars are *admission priority* without unblocking, solid bars are *discharge priority*, and grey bars are *admission priority* with unblocking.

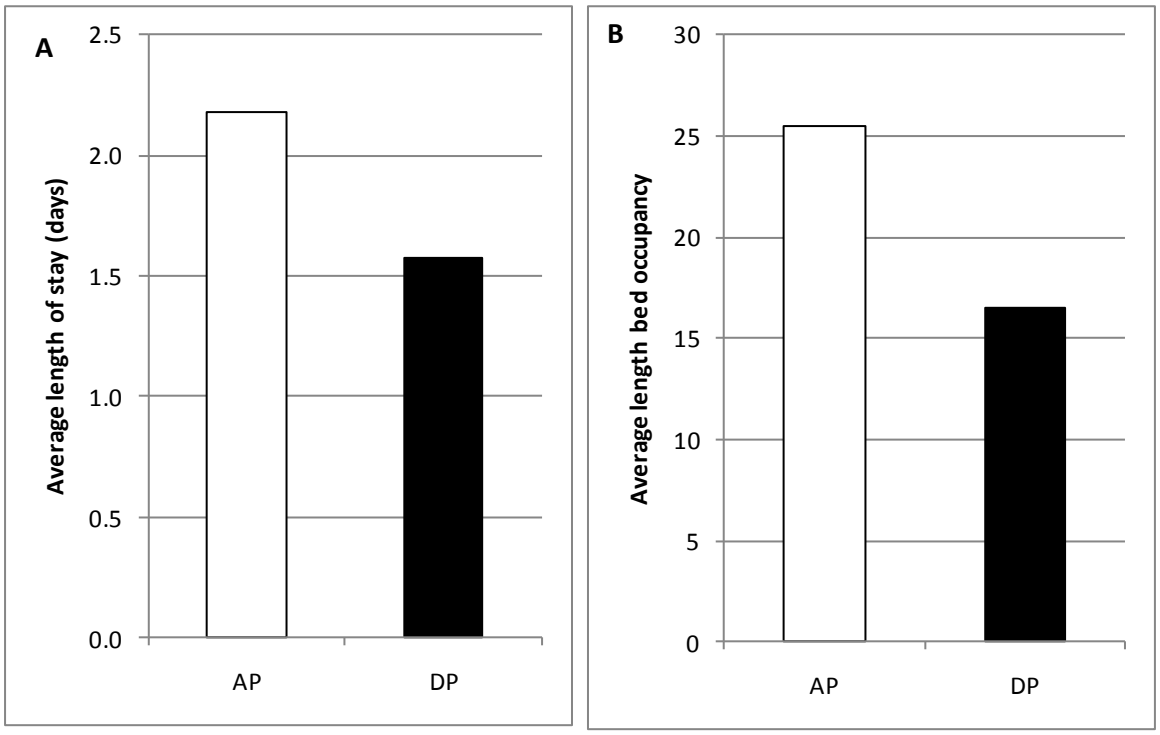

Fig. 5. The average length of stay (A) and the average bed occupancy (B) in a 40 bed model when patient arrival is set to 12 patients arriving at the start of each day (enough to occupy ~95% of system capacity). Open bars show results when *admission priority* is used, and the solid bars when *discharge priority* is used.

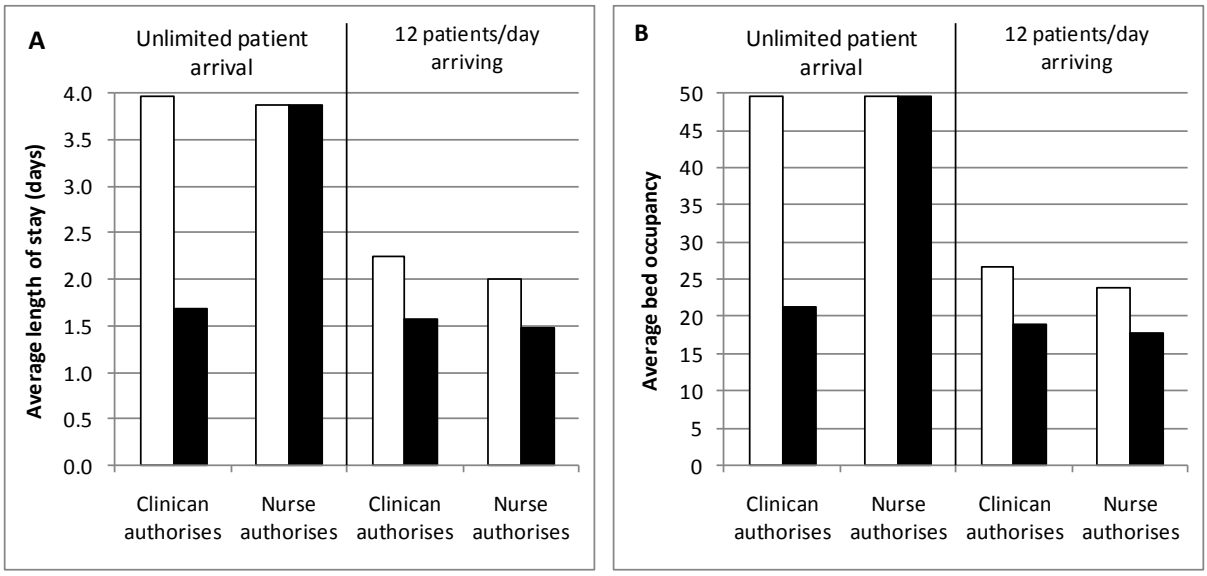

Figure 6. Average length of stay (A) and average bed occupancy (B) in a 50 bed model where patient entry into the system is permitted by either the doctor (the limiting resource) or the nurse (a non-limiting recourse). Open bars show results when *admission priority* is used, and the solid bars when *discharge priority* is used. The left panel of each graph shows results when an unlimited pool of patients is available for entry into the system, and the right panel shows results when patient arrival is fixed at 12 patients/ day (enough to occupy ~95% of system capacity).

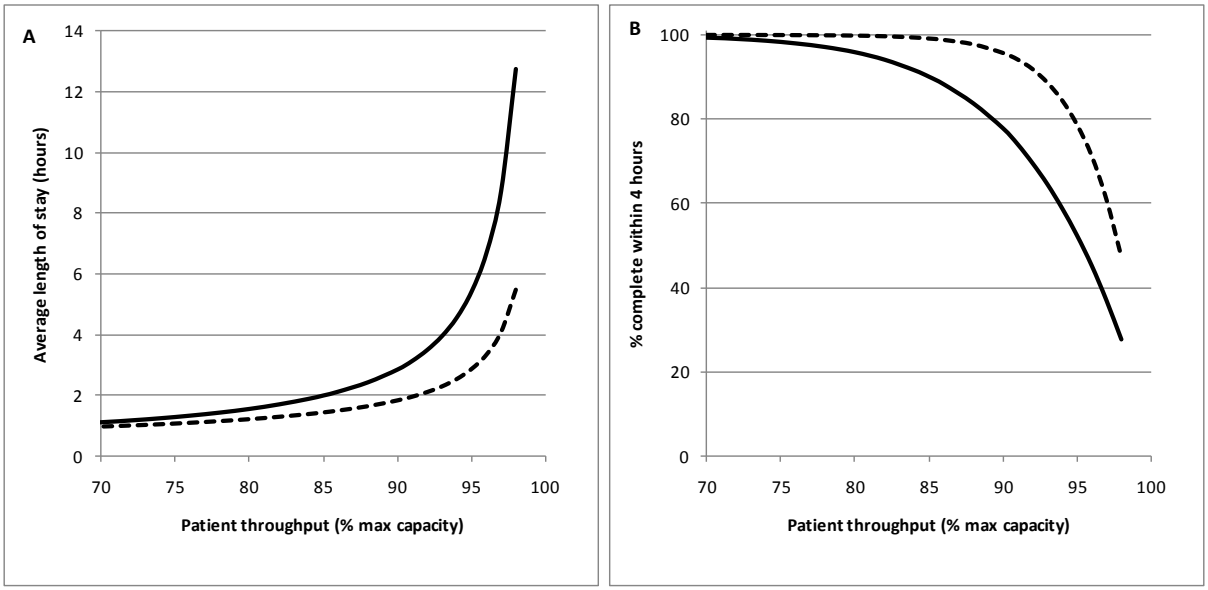

Fig.7. The relationship between patient arrival rate (expressed as a percentage of maximum system capacity) and average length of stay (A) and percentage of patients discharged within 4 hours (B). The solid line shows results when *admission priority* is used, and the dashed line when *discharge priority* is used.

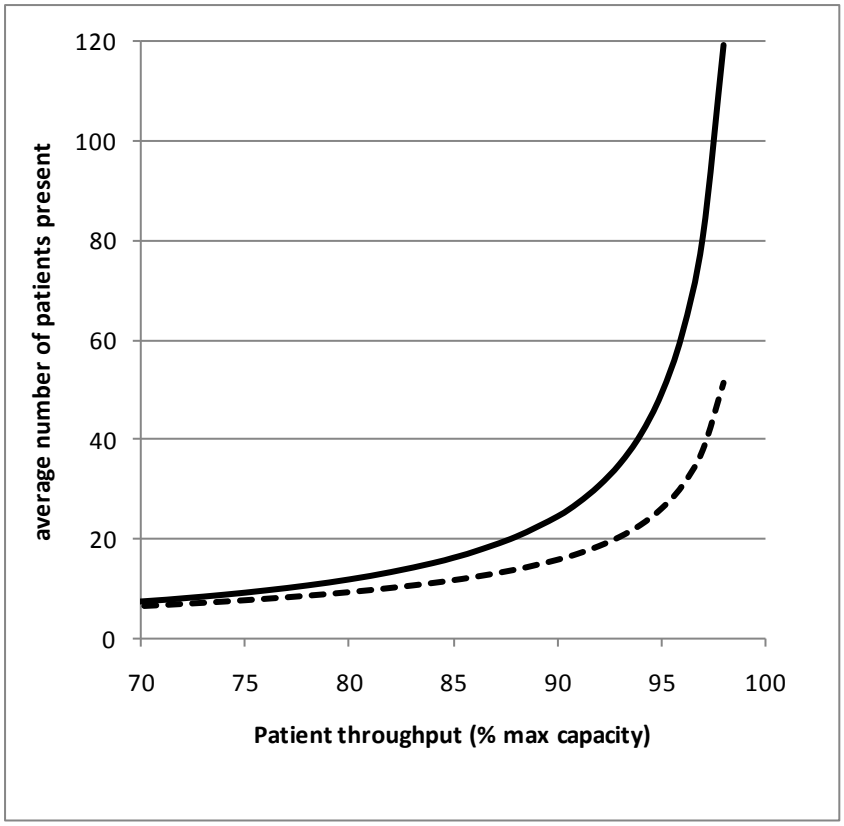

Fig.8. The relationship between patient arrival rate (expressed as a percentage of maximum system capacity) and average occupancy of unit . The solid line shows results when *admission priority* is used, and the dashed line when *discharge priority* is used.

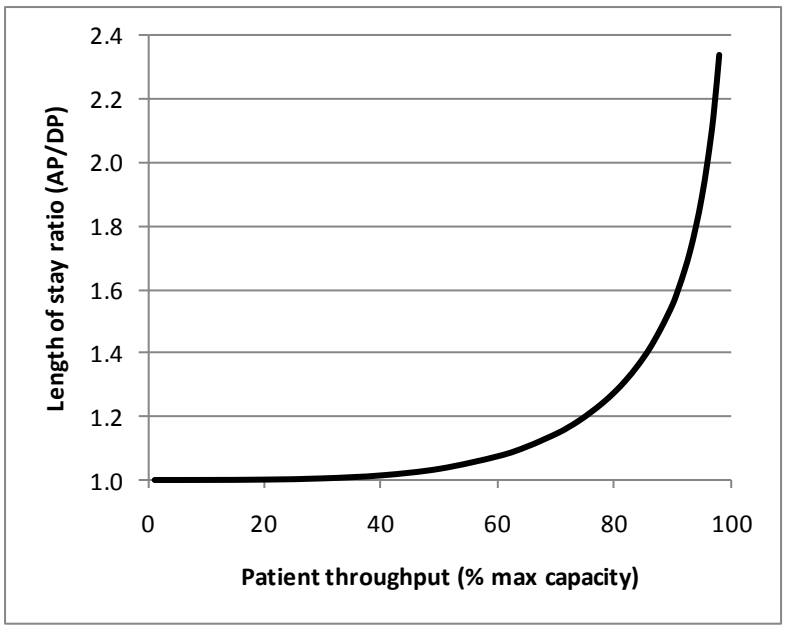

Fig.9. The relative length of stay between using *admission priority* (AP) and *discharge priority* (DP) at varying patient arrival rates (expressed as a percentage of system capacity).

WHAT IS ALREADY KNOWN ON THIS TOPIC
• Queuing in any system increases as resource utilisation increases.
WHAT THIS STUDY ADDS
• Consistently prioritising discharge, and then reducing priority back to admission, is expected to reduce average length of stay and bed occupancy in emergency and elective scenarios. • Prioritising only the discharge activity and only when under pressure to free beds may have little impact on length of stay or patient throughput. • Increasing the number of beds when staff are already fully utilised is likely to increase length of stay and costs with no increase in patient throughput (the number of patients treated may actually reduce as time is spent monitoring the increased number of patients within the hospital).